CLASSIFICAÇÃO DE CIBERATAQUES COM XAI EM SISTEMAS DE ABASTECIMENTO DE ÁGUA: UMA APLICAÇÃO COM MLP E SHAP

Rafael Barreto[1]; Jordana Alaggio[2], Gustavo Meirelles[3] & Bruno Brentan[4]

Palavras-Chave –Sistemas de Abastecimento de Água; Segurança Cibernética, Inteligência Artificial Explicativa

INTRODUÇÃO

A digitalização dos sistemas de abastecimento de água (SAA) vem promovendo avanços significativos em eficiência operacional, automação e resiliência. Contudo, essa crescente conectividade traz também vulnerabilidades importantes, tornando os SAA alvos potenciais de ciberataques com impactos severos sobre o fornecimento de água, a saúde pública e a confiança social (Taormina et al., 2017). Em um cenário onde os ataques cibernéticos a infraestruturas críticas se intensificam, é fundamental o desenvolvimento de soluções inteligentes para detecção, classificação e mitigação dessas ameaças.

Nesse contexto, técnicas de aprendizado de máquina têm sido empregadas com sucesso para identificar padrões anômalos nos dados operacionais provenientes de sistemas SCADA. No entanto, a principal limitação de muitos desses métodos reside em sua natureza de "caixa-preta", dificultando a interpretação dos resultados por operadores e profissionais de segurança (Samek et al., 2017).

Este trabalho propõe o uso de redes neurais artificiais do tipo Multilayer Perceptron (MLP) combinadas com técnicas de Inteligência Artificial Explicativa (XAI), mais especificamente o algoritmo SHAP (SHapley Additive exPlanations), para classificação de ciberataques com foco em interpretabilidade e priorização de variáveis relevantes. Ao integrar XAI ao processo de detecção, busca-se não apenas melhorar o desempenho do modelo, mas também oferecer explicações sobre os fatores que influenciam suas decisões, fornecendo uma base confiável para ações operacionais, manutenção preditiva e reforço de segurança cibernética em infraestruturas hídricas.

METODOLOGIA

O desenvolvimento da metodologia proposta envolveu quatro etapas principais: (i) preparação da base de dados BATADAL, (ii) construção do modelo de classificação baseado em redes neurais Multilayer Perceptron (MLP), (iii) aplicação do algoritmo SHAP para explicabilidade e (iv) análise de redução de variáveis com reavaliação do desempenho.

1. Preparação da Base de Dados

A base de dados utilizada é proveniente do desafio internacional BATADAL (Battle of the Attack Detection Algorithms), que fornece séries temporais simuladas de operação hidráulica e elétrica do sistema fictício C-Town. Essa rede possui reservatórios, tanques, bombas, válvulas e sensores de vazão e pressão distribuídos por diversos pontos. Os dados abrangem variáveis contínuas (níveis, vazões, pressões) e categóricas (status de bombas e válvulas) em intervalos de tempo regulares, com rotulação binária para cada instante: normal ou sob ataque.

[1]) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte/MG, rafaelb.ferreira@outlook.com
[2]) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte/MG, jordanaalaggio@gmail.com
[3]) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte/MG, limameirelles@gmail.com
[4]) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte/MG, brentan@ehr.ufmg.br

2. Construção do Modelo MLP

A arquitetura do modelo de classificação consistiu em uma rede MLP com múltiplas camadas ocultas, utilizando funções de ativação ReLU e algoritmo de retropropagação com otimização via Adam. O conjunto de entrada inicial contemplou as 43 variáveis originais disponíveis na base de dados. O treinamento foi realizado com dados rotulados, utilizando validação cruzada e divisão em conjuntos de treino (80%) e teste (20%), com controle de overfitting via early stopping.

Para avaliação de desempenho foi a calculada a matriz de confusão dos resultados e suas respectivas métricas: sensibilidade, especificidade, precisão e acurácia. Adicionalmente foi calculado um score que representa a média da sensibilidade e da especificidade.

3. Aplicação do Algoritmo SHAP

Após o treinamento do modelo MLP, foi aplicada a técnica SHAP (SHapley Additive exPlanations) para análise de interpretabilidade. O SHAP permite identifica e quantifica a contribuição de cada variável de entrada nas previsões realizadas pelo modelo, permitindo compreender os fatores que influenciam o comportamento da rede neural.

4. Seleção de Variáveis e Novo Modelo

A partir dos valores SHAP, foi estabelecido um limiar de importância mínima para filtrar as variáveis menos relevantes. Com isso, criou-se dois modelos MLP com subconjuntos das variáveis de entrada: Um modelo treinado com as variáveis de maior importância para a previsão e um modelo com as variáveis de menor importância para a previsão, segundo SHAP.

RESULTADOS

Os resultados demonstraram que o modelo MLP treinado apenas com as variáveis de maior importância para a previsão do status da rede alcançou altas taxas de sensibilidade, especificidade, precisão e acurácia na classificação dos ataques. As variáveis mais influentes identificadas pelo SHAP foram principalmente associadas aos sensores de vazão de bombas e válvulas críticas (F_PU7, F_PU8, F_V2), refletindo o impacto direto das ações de ataque nessas estruturas.

A remoção de variáveis com alta importância SHAP afetou na qualidade do modelo produzindo resultados com mais falsos positivos e, portanto, resultados menos confiáveis.

REFERÊNCIAS

SAMEK, W.; WIEDEMANN, S.; MÜLLER, K. R. *Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models*. ITU Journal: ICT Discoveries, Geneva, v. 1, n. 1, p. 39–48, 2017. Disponível em: https://arxiv.org/abs/1708.08296. Acesso em: 29 maio 2025.

TAORMINA, R. et al. *Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks*. Journal of Water Resources Planning and Management, Reston, v. 144, n. 8, 2018. DOI: https://doi.org/10.1061/(ASCE)WR.1943-5452.0000961.

AGRADECIMENTOS

Á Fapemig pelo financiamento à participação no congresso sob precoce número PCE-00429-25.